لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- پاسخ ارسالی واضح و خوانا باشد.
- در هر کدام از سوالات، اگر از منابع خاصی استفاده کرده‌اید باید آن را ذکر کنید.
- اگر با افرادی همفکری کرده‌اید، نام ایشان را ذکر کنید.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد.
- تمام پاسخ‌های خود را در یک فایل با فرمت RL_HW#_[SID]_[Fullname].zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می توانید تا سقف ۲ روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید و در مجموع ۵ روز تاخیر مجاز برای تمارین در اختیار دارید.

سوال ۱: یادگیری تقویتی برون خط

Conservative Q-Learning یک روش یادگیری تقویتی برون خط است که تلاش می کند محافظه کارانه ترین تابع Q را بر اساس داده ها پیدا کند.

$$\begin{aligned}\text{CQL} = \min_{Q}\max_{\mu}\, \alpha\left(\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\mu(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})] - \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\hat{\pi}_\beta(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})]\right)\\ + \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q(\mathbf{s},\mathbf{a}) - \hat{\mathcal{B}}^{\pi_k}\hat{Q}^k(\mathbf{s},\mathbf{a})\right)^2\right] + \mathcal{R}(\mu)\end{aligned}$$

(آ) هدف از عبارت اول چیست؟

$$\mathbb{E}_{\mu}[Q] - \mathbb{E}_{\hat{\pi}_\beta}[Q]$$

(ب) نشان دهید اگر $\mathcal{R}(\mu) = H(\mu)$ داریم،

$$\text{CQL} = \min_{Q}\alpha\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[\log\sum_{\mathbf{a}}\exp(Q(\mathbf{s},\mathbf{a})) - \mathbb{E}_{\mathbf{a}\sim\hat{\pi}_\beta(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})]\right] + \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q - \hat{\mathcal{B}}^{\pi_k}\hat{Q}^k\right)^2\right]$$

توضیح دهید که این نرمالسازی چگونه هدف محافظه کارانه را تعدیل میکند.

(ج) فرض کنید $\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\hat{\pi}_\beta(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})]$ در رابطه CQL وجود ندارد. در این صورت با مشتق گیری رابطه ی زیر را برای $\hat{Q}^{k+1}$ بدست آورید،

$$\forall \mathbf{s},\mathbf{a}\in\mathcal{D}, k,\quad \hat{Q}^{k+1}(\mathbf{s},\mathbf{a}) = \hat{\mathcal{B}}^{\pi}\hat{Q}^k(\mathbf{s},\mathbf{a}) - \alpha\frac{\mu^*(\mathbf{a}\mid\mathbf{s})}{\hat{\pi}_\beta(\mathbf{a}\mid\mathbf{s})}$$

که

$$\mu^* = \text{argmax}_{\mu}\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\mu(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})] + \mathcal{R}(\mu)$$

(د) فرض کنید با احتمال حداقل $1-\delta$ یک کران برای تخمین Bellman Operation داریم،

$$\forall Q,\mathbf{s},\mathbf{a}\in\mathcal{D}, \left|\hat{\mathcal{B}}^{\pi}Q(\mathbf{s},\mathbf{a}) - \mathcal{B}^{\pi}Q(\mathbf{s},\mathbf{a})\right| \le C_\delta(\mathbf{s},\mathbf{a}).$$

با فرض قسمت قبل اثبات کنید که برای نقطه ثابت Q در CQL کران بالای زیر وجود دارد.

$$\hat{Q}^{\pi}(\mathbf{s},\mathbf{a}) \le Q^{\pi}(\mathbf{s},\mathbf{a}) - \alpha\left[(I-\gamma P^{\pi})^{-1}\left[\frac{\mu}{\hat{\pi}_\beta}\right]\right](\mathbf{s},\mathbf{a}) + \left[(I-\gamma P^{\pi})^{-1}C_\delta\right](\mathbf{s},\mathbf{a})$$

کران بیان شده یک کران بالا برای تابع Q بدست آمده توسط CQL بدون در نظر گرفتن عبارت $\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\hat{\pi}_\beta(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})]$ است.

سوال ۲: یادگیری تقویتی وارونه

در یادگیری تقویتی وارونه هدف پیدا کردن سیاست و تابع پاداشی است که منجر به خط سیر (trajectory) های مشابه با دیتای جمع آوری شده از خبره باشد. یکی از این روش ها جور کردن امیدریاضی ویژگی های مربوط به حطوط سیر است.
همچنین یادگیری تقویتی وارونه را به صورت یک بازی میان یک عامل سیاست و یک عامل تعیین پاداش بیان کرد که عامل سیاست تلاش دارد نسبت به عامل خبری بیشتری پاداش را دریافت کند و عامل تعیین پاداش به طور خصمانه تلاش در تعیین تابع پاداش به صورتی دارد که عکس این اتفاق رخ دهد،

$$\min_{\pi\in\Pi}\max_{f\in\mathcal{F}_r} J(\pi_E, f) - J(\pi, f). \tag{۱}$$

این فرمول بندی از یادگیری تقویتی وارونه توسط روش Generative Adversarial Imitation Learning (GAIL) پیاده سازی شده است.

(آ) برتری روش یادگیری تقویتی وارونه نسبت به روش تقلید رفتار (Behavior Cloning) چیست؟

(ب) یکی از مسائل موجود در روش های مبتنی بر تطبیق ویژگی ابهام است. روش Maximum Entropy Inverse Reincforcement Learning را توضیح دهید و بیان کنید که چگونه این چالش را پر طرف کرده است.

(ج) تابع هدف GAIL به صورت زیر است.

$$\min_{\pi}\max_{D} \mathbb{E}_{\pi}[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1-D(s,a))] - \lambda H(\pi)$$

توضیح دهید GAIL چگونه بازی min-max عبارت ۱ را حل می کند و نقش D و H در آن چیست؟

(د) روش های تطبیق ویژگی مانند MaxEnt IRL چه مشکلی دارند و این مشکل چگونه در GAIL برطرف می شود؟

سوال ۳: یادگیری تقویتی مبتنی بر مدل

یکی از روش های یادگیری تقویتی مبتنی بر مدل Model-Based Policy Optimization (MBPO) است که در آن به صورت ترکیبی از مدل محیط و تعامل با محیط واقعی استفاده می شود.
برای یادگیری تقویتی برون خط روش های دیگری همچون Model-based Offline Reinforcement Learning(MOReL) و Conservative Offline Model-Based Policy Optimization (COMBO) برای استفاده از داده های برون خط و بدون تعامل با محیط ارائه شده اند.

(آ) توضیح دهید که روش هایی که تنها از مدل محیط استفاده می کنند و روش هایی که کاملا متکی به تعامل با محیط هستند چه مشکلی دارند و روش های ترکیبی چگونه به نتایج بهتری دست پیدا می کنند؟

(ب) توضیح دهید MBPO چگونه این ایده را در بهینه سازی سیاست دخیل می کند و trade-off انتخاب مقادیر مختلف برای طول rollout های مبتنی بر مدل به چه صورت است؟

(ج) علت اینکه از MBPO نمیتوان در حالت برون خط استفاده کرد چیست؟

(د) در روش MOReL تردید (uncertainty) مدل در مورد جفت حالت_عمل های مختلف سنجیده می شود و با تغییر MDP بر این اساس و تعریف یک MDP جدید از رفتن به این حالات جلوگیری می شود. یک روش برای تعیین میزان تردید ارائه دهید.

(ه) در روش COMBO از استراتژی محافظه کارانه برای حل مسئله یادگیری تقویتی برون خط بر اساس مدل استفاده می شود. توضیح دهید COMBO چگونه ایده های دو روش CQL و Dyna را تلفیق می کند و به هدف هر دو روش دست پیدا می کند.

(و) روش های محافظه کارانه و برساس تردید را با هم مقایسه کنید و نقاط قوت و ضعف هر کدام را توضیح دهید.

سوال ۴: یادگیری با بازخورد جزئی Bandit Learning

یکی از الگوریتم های شناخته شده در Multi-Armed Bandit روش Upper Confidence Bound(UCB) است. فرض کنید عامل در گام t سعی در انتخاب یک عمل مناسب دارد. و تا الان $T_i(t-1)$ نمونه از عمل i انجام داده است و نتیجه آن ها را دیده است و به طور میانگین پاداش $\hat{\mu}_i(t-1)$ از آن ها دریافت کرده است. در این صورت UCB با اطمینان δ به این صورت تعریف می شود.

$$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} \infty & \text{if } T_i(t-1)=0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \text{otherwise .} \end{cases}$$

در ادامه فرض کنید پاداش هر عمل یک متغیر تصادفی محدود به بازه $[-1,1]$ است.

(آ) با توجه به نابرابری به کمک نابرابری Hoefding[1] توضیح دهید که UCB چگونه یک کران بالا برای میانگین پاداش هر عمل تعیین می کند.

(ب) نشان دهید

$$\mathbb{E}\left[T_i(n)\right] \leq 3 + \frac{16\log(n)}{\Delta_i^2}.$$

که $\Delta_i = R_{\max} - R_i$ برابر با میانگین regret مربوط به عمل i است.

(ج) نشان دهید اگر قرار دهیم $\delta = \frac{1}{n^2}$

$$R_n \leq 3\sum_{i=1}^{k}\Delta_i + \sum_{i:\delta_i>0}\frac{16\log\left(n\right)}{\Delta_i}$$

(د) حال با همان مقدار قبلی برای δ نشان دهید

$$R_n \leq 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k}\Delta_i.$$

[1] https://en.wikipedia.org/wiki/Hoeffding%27s_inequality